**Source**: **TBD/Editor[[1]](#footnote-2)\***

**Title: IVAS Permanent Document EVS-8a: Test plans for selection phase including lab task specification**

**Version: v.0.0.1**

**Agenda Item: 7.5**

**Document History:**

|  |  |  |
| --- | --- | --- |
| v.0.0.1 | 24 May 2021 | Initial Skeleton |

**Reference Input Documents:**

|  |  |
| --- | --- |
| S4-210836 | On reference designs for IVAS codec tests |

**[**

1. Introduction

This document contains the set of test plans for the Selection Phase of the Codec (IVAS).

2. References, Conventions, and Contacts

## 2.1 Permanent Documents

## 2.2 Reference Documents

## 2.3 Key Acronyms

## 2.4 Contact Names

3 Roles and Responsibilities

## 3.1 Overview of the Selection Test Process

## 3.2 Allocation of Additional Roles

## 3.3 Responsibilities

3.3.1 Proponent Companies

3.3.2 Listening Laboratories

3.3.3 Host Laboratory

3.3.4 Cross-check Laboratory

3.3.5 Global Analysis Laboratory

4 Information relevant to all Experiments

## 4.1 General Technical Notes

## 4.2 General Consideration of Experiments

## 4.3 Opinion Scales

## 4.4 Material

4.4.1 Speech Material

4.4.2 Noise Material

4.4.3 Captured Music and Mixed Content Material

## 4.5 Listening Systems and Listening Environments

## 4.6 Experimental Procedure

## 4.7 Results and Analysis

5. Subjective Experiments

Annex A: Sample Instructions to Subjects and Data Collection

Annex B: Presentation Orders

Annex C: Data to be Provided by LL

Annex D: Obligations and Task for the Listening Laboratories

Annex E: Host Laboratory Tasks

## E.1 Included tasks

## E.2 Excluded tasks

Annex F: Cross check Laboratory Tasks

## F.1 Included tasks

## F.2 Excluded tasks

Annex G: GAL Tasks

## G.1 Tasks

## G.2 Statistical analysis of results

Annex H: Selection Testing Timeline

]

**Appendix I: Examples of test designs potentially relevant for IVAS codec testing**

## Introduction

This Appendix contains a collection of experimental designs that are deemed potentially relevant for IVAS codec testing. When creating the IVAS codec selection and characterizations test plans SA4 may decide to resort to concepts of these designs.

## Example P.800 DCR test of spatial (FoA) speech [S4-210836]

## Test Purpose

Build an opinion about suitability of modified P.800 DCR test methodology for quality assessments of immersive conversational speech.

## Test Outline

* 2 Experiments
  + Exp1: use case ‘immersive conferencing’ with Ambisonics (FoA) spatial speech, 6 content type categories constructed as follows:
    - Model-based relying on convolution of raw mono clean speech sentences convolved with (FoA) Spatial Room Impulse Responses respective various talker positions relative to a capture point. The Spatial Room Impulse Responses were recorded in the respective conference rooms.
    - Spatialized sentences are combined to sentence pairs and mixed with spatial (FoA) ambient noise.
    - 2 relatively low background noise levels (30, 40 dB SNR, based on level normalization according to ITU-R BS.1770-4 [4])
    - Reverberance typical for 2 conference rooms (large and small)
    - 2 talker interactions types: sentence pairs with and without ‘overtalking’ (1s overtalk)
    - Language: Polish
    - Lab: Dolby Wroclaw (Poland)
  + Exp2: Immersive telephony while on the move (outside) with Ambisonics (FoA) spatial speech, 6 content type categories constructed as follows:
    - Model-based relying on convolution of raw mono clean speech sentences convolved with (FoA) Spatial Room Impulse Responses respective various talker positions relative to a capture point. The Spatial Room Impulse Responses were recorded in the respective test environments (car) or a low-echoic room approximating the other environments.
    - Spatialized sentences are combined to sentence pairs and mixed with spatial (FoA) ambient noise.
    - Moderate to high background noise levels (15, 20, 25dB SNR, based on level normalization according to ITU-R BS.1770-4 [4])
    - Various environments: street, car, public indoor (shopping mall, subway station)
    - No talker interactions (no ‘overtalking’): sentence pairs without ‘overtalking’ (1s gap)
    - Language: American English
    - Lab: Dolby San Francisco (USA)/remote (home environment)

## General Consideration of Experiments

* Six categories of content types.
* 30 subjects, five listening panels (six subjects per panel), each panel with an independent randomization.
* Five samples per category (one for each listening panel).
* Randomizations constructed under “partially-balanced/randomized blocks” experimental design described in “Practical procedures for subjective testing”, [5].
* Every condition has 30 different samples passed through it (6 categories x 5 panels). Each of these are voted on by the 6 subjects in the panel, giving: (30 samples x 6 subjects/panel) = 180 (150) votes per condition.
* 30 test conditions x 6 categories = 180 DCR trials.
* Average trial duration: 16 s (6.5 s reference sample +0.5 s silence + 6.5 s test sample + 2.5 s voting period).
* Test duration: ~1.6 h per listening panel. Test duration comprises 50% of actual listening/voting time (48 min) and 50% test overhead including orientation, instructions, preliminaries, and rest breaks
* The listening sessions were split into a number of sub-sessions with breaks in between to allow for the subject to relax. This was to avoid listener fatigue.
* Test platform: Dolby-internal

## Degradation references (anchors)

According to ITU-T Rec. P.811 Appendix II, P.811 [6] overall quality scores strongly correlate with P.800 DCR scores if the latter is run with modified instructions and degradation references that span both signal and spatial quality dimensions. P.811 suggests using P.50 MNRU for signal degradation anchors and SDRU/ESDRU for spatial degradation anchors. P.50 MNRU is a modulated noise reference unit with P.50-artificial voice weighting. SDRU/ESDRU are spatial degradation reference units defined for stereo signals that gradually, depending on a degradation parameter α, impair the stereo image without substantially causing signal distortions. A random process additionally introduces temporal fluctuations ranging from the original to the maximally degraded stereo image. The ESDRU applies a more sophisticated random process.

We followed this recommendation and adapted the P.50 MNRU and the ESDRU to derive degradation anchors for our P.800 experiments with binauralized FOA content.

For the P.50 MNRU the adaptation is that it is coherently applied (same seed) to all 4 FOA signals. This has the perceptual effect that the spatial direction of the introduced signal distortion coincides with the spatial signal direction. Thus, the introduced signal distortion does not significantly affect the spatial image.

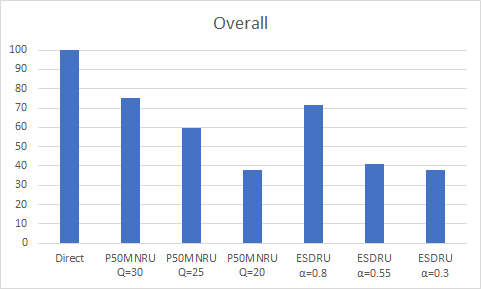
The ESDRU on the other hand is directly applied to the two binaural channels after binaural rendering of the FOA signal.

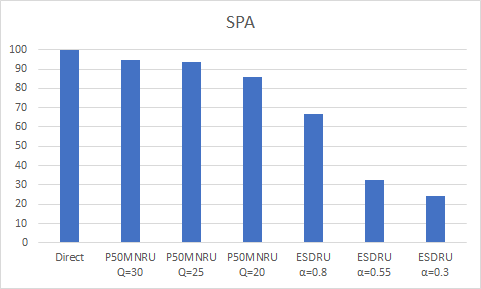
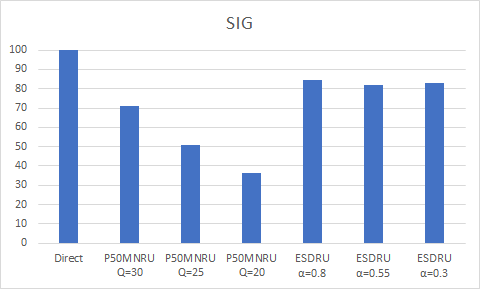
A limited subjective experiment was carried out to

1. verify the suitability of these degradation anchors,
2. to verify the basic assumption that the P.50 MNRU has little impact on spatial distortion and vice-versa that the ESDRU has little impact on perceived signal distortion, and
3. to find suitable P.50 MNRU and ESDRU degradation parameters Q and, respectively, α.

In the experiment 6 FOA voice vectors were degraded either with P.50 MNRU values of Q=30, 25, and 20 dB or with ESDRU parameter values of α = 0.8, 0.55, and 0.3. These vectors were evaluated in a Mushra test (with 3 expert listeners) with the three quality attributes *overall quality (Overall)*, *signal quality (SIG)*, and *spatial quality (SPA).*

The results are displayed in the following plots:





From the plots, the following observations can be made:

* The P.50 MNRU degradation affects mainly signal (SIG) and Overall quality while spatial quality (SPA) is less impacted.
* The ESDRU degradation affects mainly spatial (SPA) and Overall quality while signal quality (SIG) is less impacted.
* The P.50 MNRU induced signal degradation appears a bit too strong and should be softened for the P.800 tests.
* The ESDRU induced degradation is too strong, which results in that spatial and overall quality start to saturate at the lower end. Consequently, for the P.800 tests, it was decided to increase the α parameters.

## Factors and conditions

## Main experiment

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 11 | Dolby-internal FoA coding system |
|  |  |  |
| **Codec references** |  |  |
| Codec references | 12 | Multi-mono 4xEVS operated at  4\*8, 4\*9.6, 4\*13.2, 4\*16.4, 4\*24.4, 4\*32, 4\*48, 4\*64, 4\*96 kbps with DTX off and  4\*13.2, 4\*16.4, 4\*24.4 kbps with DTX on |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | Nominal input level |
| P.50 MNRU (applied to all FoA components) | 3 | Q=22, 27, 32 dB (all: nominal level) |
| ESDRU [6] | 3 | α = 0.55, 0.7, 0.85 (output loudness forced to nominal level) |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FoA) Spatial Room Impulse Responses respective various talker positions relative to a capture point and spatial (FoA) ambient noise mixing |
| Binaural renderer | 1 | FoA to binaural rendering according to [7] |
| Audio sampling frequency/bandwidth | 2 | 48 kHz/SWB except for 4xEVS@4\*8kbps which is 48 kHz/WB |
| Content types (categories) | 6 | Exp1: 6 Different conference rooms and talker interactions  Exp2: 6 Different background noise types and levels |
| Kind of samples | 1 | Sentence pair uttered by different talkers and genders (3 male and 3 female) |
| Number of samples | 5 | per content type |
| Input frequency mask | 1 | Flat |
| Nominal output loudness | 1 | -26 LKFS (ITU-R BS.1770-4 [4]) |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations | 5 | 5 panels of 6 listeners |
| Rating Scale | 1 | DCR with modified instructions |
| Replications | 1 |  |
| Languages | 1 | Exp1: Polish, Exp2: American English |
| Listening System | 1 | High-quality headphone for diotic presentation |
| Listening Environment | 1 | No room noise |

## Preliminaries (familiarization of listeners)

|  |  |  |
| --- | --- | --- |
| **Main Codec Conditions** |  |  |
| Codec under Test (CuT) | 0 |  |
| Codec references | 5 | Multi-mono 4xEVS operated at  4\*8, 4\*13.2, 4\*24.4, 4\*48, 4\*64, with DTX off |
|  |  |  |
| **Other references** |  |  |
| Direct | 1 | Nominal input level |
| P.50 MNRU (applied to all FoA components) | 3 | Q=22, 27, 32 dB (all: nominal level) |
| ESDRU [6] | 3 | α = 0.55, 0.7, 0.85 (output loudness forced to nominal level) |
|  |  |  |
| **Common Conditions** |  |  |
| Test item generation: pre-processing incl. spatialization | 1 | Model-based relying on convolution of raw mono clean speech sentences convolved with (FoA) Spatial Room Impulse Responses respective various talker positions relative to a capture point and spatial (FoA) ambient noise mixing |
| Audio sampling frequency/bandwidth | 1 | 48 kHz/SWB except for 4xEVS@4\*8kbps which is 48 kHz/WB |
| Content types (categories) | 6 | Exp1: 6 Different conference rooms and talker interactions  Exp2: 6 Different background noise types and levels |
| Number of samples | 1 | per content type |
| Input frequency mask | 1 | Flat |
| Nominal output loudness | 1 | -26 LKFS (ITU-R BS.1770-4 [4]) |
| Listening Level | 1 | 73 dB SPL |
| Listeners | 30 | Naïve Listeners |
| Randomizations | 1 | Same randomization for the 5 panels of 6 listeners |
| Rating Scale | 1 | DCR with modified instructions |
| Replications | 1 |  |
| Languages | 1 | Exp1: Polish, Exp2: American English |
| Listening System | 1 | High-quality headphone for diotic presentation |
| Listening Environment | 1 | No room noise |

## Instructions to listeners and Degradation Scale

The following presents the modified DCR test instructions given to the subjects and the five-point degradation category scale used in the test:

**"Evaluation of the quality of future 3D audio telephony and conferencing systems"**

In this experiment you will hear pairs of speech samples that have been recorded through various experimental 3D audio telephone and conferencing equipment. You will listen to these samples through a set of stereo headphones.

What you will hear is a first sample containing one pair of sentences from two talkers, a short period of silence, and a second sample. You will evaluate the OVERALL quality of the second sample compared to the quality of the first sample.

You should listen carefully to each pair of samples. As soon as a sample pair has been completely played back, you should register your opinion on ANY kind of degradation of the second sample compared to the first sample. Please consider in your vote, besides, e.g., the quality of the speech or other sounds, also any change in the perceived location of voices or sounds or changes in spatial width.

Then, when the system requests your vote, please record your opinion on the OVERALL quality using the following scale:

The OVERALL quality DEGRADATION of the Second Compared to the First is:

5: Inaudible

4: Audible but not annoying

3: Slightly annoying

2: Annoying

1: Very annoying

You will have five seconds to record your answer by pushing the button corresponding to your choice. There will be a short pause before the presentation of next pair of sentences.

We will begin with a short practice session to familiarize you with the test procedure. The actual tests will take place during multiple sessions with short breaks in between.

# Degradation Scale

The OVERALL quality DEGRADATION of the Second Compared to the First is:

5: Inaudible

4: Audible but not annoying

3: Slightly annoying

2: Annoying

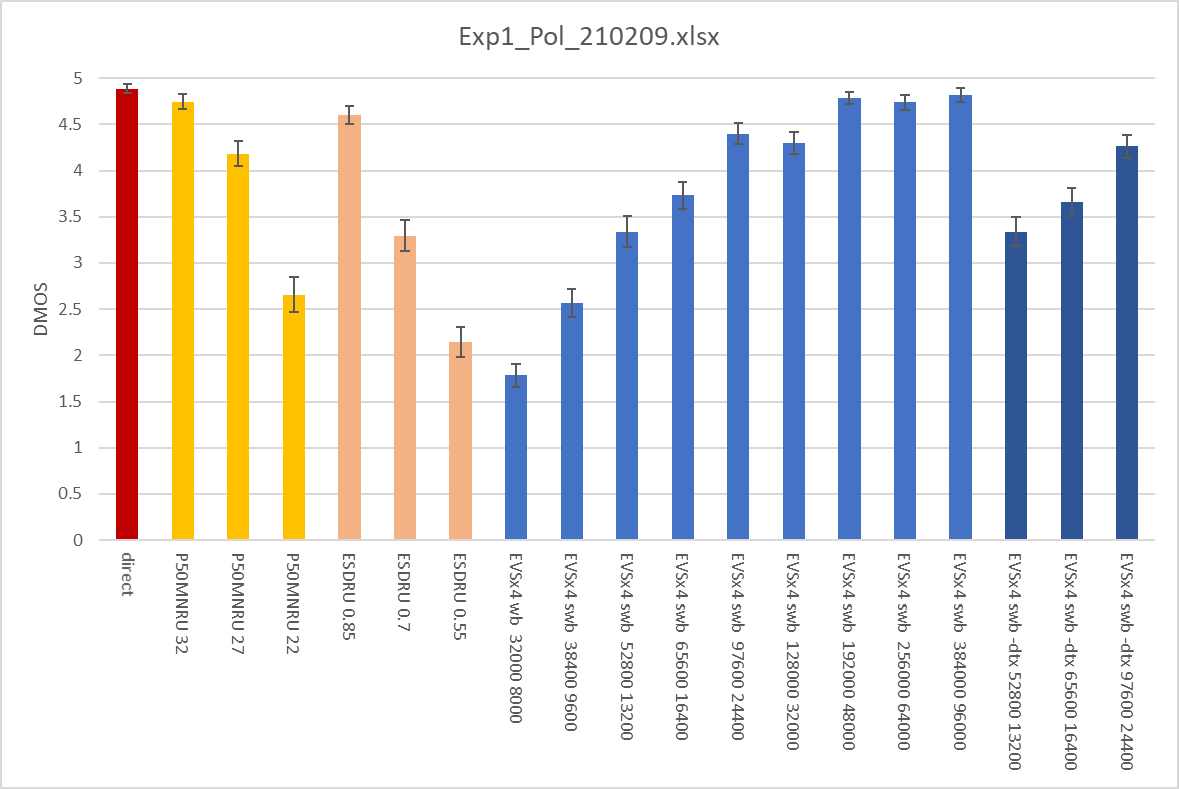
1: Very annoying

## Results

## Exp1: Use case ‘immersive conferencing’ with FoA speech (Polish)

Below is a graph showing the MOS scores with 95%CI of the direct, the degradation reference and the EVS multi-mono conditions. The results of the CuT are not presented as they are not relevant for the purpose of this contribution.

## Degradation Mean Opinion Scores



## Observations:

The mean opinions scores observed for the reference systems and especially the EVS reference system is very consistent with the expectations that build on the performance characterization of that codec [8]. This is an indicator of the good resolution of the test. At high bit rates there is some non-surprising saturation effect.

## Exp2: Immersive telephony while on the move (outside) with FoA audio (American English)

**Editor’s note:** This experiment was still not finalized when the contribution was submitted. SA4 may add the results when available.

1. \* TBD [↑](#footnote-ref-2)